Complex Patterns of Association between Pleiotropy and Transcription Factor Evolution

Kevin N. Chesmore, Jacquelaine Bartlett[†], Chao Cheng, and Scott M. Williams*,[†]

Department of Genetics, Geisel School of Medicine, Dartmouth College, Hanover, NH

[†]Present address: Departments of Epidemiology and Biostatistics and Genetics and Genome Sciences, Case Western Reserve University, Cleveland, OH

*Corresponding author: E-mail: smw154@case.edu.

Accepted: September 9, 2016

Abstract

Pleiotropy has been claimed to constrain gene evolution but specific mechanisms and extent of these constraints have been difficult to demonstrate. The expansion of molecular data makes it possible to investigate these pleiotropic effects. Few classes of genes have been characterized as intensely as human transcription factors (TFs). We therefore analyzed the evolutionary rates of full TF proteins, along with their DNA binding domains and protein-protein interacting domains (PID) in light of the degree of pleiotropy, measured by the number of TF–TF interactions, or the number of DNA-binding targets. Data were extracted from the ENCODE Chip-Seq dataset, the String v 9.2 database, and the NHGRI GWAS catalog. Evolutionary rates of proteins and domains were calculated using the PAML CodeML package. Our analysis shows that the numbers of TF-TF interactions and DNA binding targets associated with constrained gene evolution; however, the constraint caused by the number of DNA binding targets was restricted to the DNA binding domains, whereas the number of TF-TF interactions constrained the full protein and did so more strongly. Additionally, we found a positive correlation between the number of protein–PIDs and the evolutionary rates of the protein–PIDs. These findings show that not only does pleiotropy associate with constrained protein evolution but the constraint differs by domain function. Finally, we show that GWAS associated TF genes are more highly pleiotropic. The GWAS data illustrates that mutations in highly pleiotropic genes are more likely to be associated with disease phenotypes.

Key words: pleiotropy, evolutionary constraint, transcription factors, ENCODE.

Introduction

Pleiotropy, a term first coined in 1910, describes the phenomenon where a single gene has multiple biological functions (Plate 1910; Stearns 2010) . Initially, pleiotropy was considered to be rare, as it was thought that most genes only possessed a single function (Plate 1910), and this idea of a single gene, single function remained throughout most of the 20th century (Stearns 2010). However, as our understanding of molecular biology has improved, it is becoming evident that pleiotropy is nearly ubiquitous (Stearns 2010), and examples were found in diverse fields of biology, ranging from normal development and aging to complex diseases (Sivakumaran et al. 2011; Wagner and Zhang 2011). Pleiotropy has been previously classified into seven types (Hodgkin 1998): 1) Artefactual—mutations affect multiple independent genes, 2) Secondary—proteins affect one biochemical process which results in a complex set of phenotypes, 3) Adoptive—proteins having different tissue specific functions, 4) Parsimonious—proteins preforming the same function in multiple pathways, 5) Opportunistic—proteins having one primary function and additional secondary roles, 6) Combinatorial—proteins having different functions depending on which proteins it is interacting with, and 7) Unifying - proteins fulfill multiple roles within a single biological pathway. A well-documented example of pleiotropy (adoptive) has been shown for *αβ-crystallin* that contributes to both cataracts and diffuse cardiomyopathy (Inagaki et al. 2006; Liu et al. 2006; Tyler et al. 2009). Another example (parsimonious) is the *CFTR* gene which results in a range of phenotypes from pancreatic insufficiency and infertility to lung infections (Vankeerberghen et al. 2002).

It was hypothesized that genes affecting multiple traits would likely experience stronger purifying selection, thus restricting the rate at which they evolve (Caspari 1952). Additionally, it was recognized that if multiple phenotypes were affected by a single mutation, not all phenotypic effects would be uniformly beneficial nor uniformly detrimental (Williams 1957). Such possibly antagonistic effects may inform our understanding of the maintenance of disease causing alleles in populations. Many studies have provided evidence for the ability of pleiotropy to constrain gene evolution, using protein–protein interaction (PPI) networks, Gene Ontology terms, and phenotypes of gene knockouts as measurements of pleiotropy (Fraser et al. 2002, 2003; Jordan et al. 2003; Fraser and Hirsh 2004; Fraser 2005; He and Zhang 2006; Kim et al. 2006; Salathe et al. 2006; Artieri et al. 2009; Chang et al. 2013). Despite the support for the existence of pleiotropic constraints, prior studies could not, for the most part, explicitly describe the underlying molecular processes that impact the rate of gene evolution and whether domain function directly affected the degree of constraint. Additional models of evolutionary constraints have also been proposed that act independently of pleiotropy (Zhang and Yang 2015), such as developmental stage (Piasecka et al. 2013), gene methylation status (Chuang and Chiang 2014), protein synthesis (Yang et al. 2010, 2014), protein folding (Drummond and Wilke 2008; Yang et al. 2010), contact density (Zhou et al. 2008), codon bias (Akashi 1994; Ran et al. 2014), and gene expression (Drummond et al. 2005; Pal et al. 2006; Gout et al. 2010; Yang et al. 2012) (currently considered the major driver of evolutionary constraint), although these have not been extensively examined in conjunction with possible pleiotropic constraints in multicellular organisms.

In recent years, the generation of high quality, large-scale, and less biased datasets, documenting gene and protein functions, have provided powerful resources to study the molecular mechanisms underlying pleiotropic constraint (Tang et al. 2014). Projects such as ENCODE (Gerstein et al. 2012; Al-Maawali et al. 2015), provide an unparalleled set of functional information on proteins within the human transcription factor (TF) network, including the number of interacting partners that serve as a measurement of pleiotropy. The breadth and quality of these and related data allow us to assess fundamental questions regarding the role of domain function in constraining gene evolution.

In this study, we looked at the number of TF–TF interactions, and the number of DNA binding targets as measures of pleiotropy. Based on the seven types of pleiotropy, these measurements of pleiotropy would primarily qualify as opportunistic, although these data likely include instances of adoptive, combinatorial and unifying pleiotropy. We explored how these two measurements of pleiotropy correlated to the evolutionary rates of the full proteins as well as each class of functional domain separately. More specifically we investigated whether the constraint caused by the two molecular functions differed with respect to how they constrained the protein's and/or domain's evolutionary rates (fig. 1A). Additionally, we were interested in how the number of TF-TF interactions and the number of DNA binding targets correlated with the number of protein interacting domains (PIDs) and DNA Binding domains (fig. 1B). We also assessed whether TFs identified in GWAS are more pleiotropic and more constrained than non-associated TFs.

Methods

Human TF Identification and Sequence Retrieval

TFs were identified as previously described (Ravasi et al. 2010); briefly, any gene that is annotated as a "TF" by the Gene Ontology database or Roach et al. (2007), as well as all genes whose Entrez description field contains the word "transcription" was considered a TF. After annotating using these automated processes, the gene list was manually curated to remove any genes that did not belong on a list of TFs, resulting in a final list of 1988 TFs in the human genome (Ravasi et al. 2010).

In order to calculate the evolutionary rates of the TF proteins and functional domains, we first collected orthologs of each of the 1988 TF genes from NCBI for the following 12 species: *Homo sapiens, Pan paniscus, Macaca mulatta, Pan troglodytes, Nomascus leucogenys, Chlorocebus sabaeus, Tarsius syrichta, Papio anubis, Callithrix jacchus, Otolemur garnettii, Saimiri boliviensis, and Gorilla gorilla.* On average we were able to identify orthologues in 9.3 species per human TF; the average of fewer than 12 orthologous sequences per TF was due to incomplete annotation and sequencing of non-human species. Each coding sequence (CDS) was identified based on the corresponding Genbank file for each gene and was later used for all evolutionary rate calculations for the full proteins as well as their functional domains.

Functional Domain Identification

To calculate the evolutionary rates of the functional domains, each human TF CDS was translated into its amino acid sequence using the translate function in Biopython package Bio.Seq (Cock et al. 2009). These protein sequences were then run through the NCBI Batch Web CD-Search Tool against the Conserved Domain Database (Marchler-Bauer et al. 2011) to identify all annotated domains in each protein. Based on the domain descriptions from the NCBI CD database, we identified all domains related to either PPIs or DNA binding. These domain sequences were then aligned to the orthlogous protein sequences for all available non-human species, using ClustalOmegaCommandline in the Biopython package (Cock et al. 2009) . This selection process allowed us to accurately define functional domains for species whose genomes are poorly annotated. All PPI domains within a single protein were concatenated to form a single sequence. Similarly, all DNA binding domains (DBD) within a single gene were concatenated. Additionally, we also counted the total number of PPI domains and DNA binding domains found in each gene.

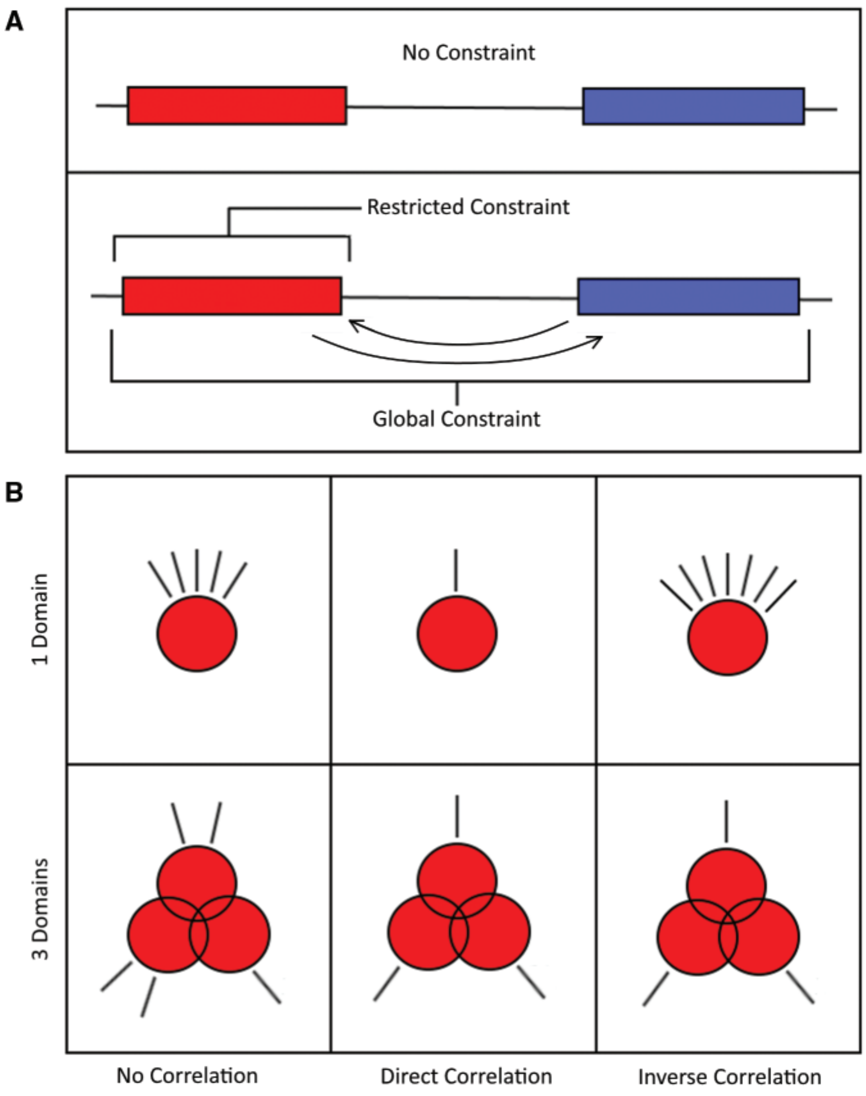

Fig. 1.—Models of constraint. (*A*) A hypothetical protein containing two domains of different molecular functions (Red or Blue). Brackets illustrate which regions of the protein are constrained, and arrows show the direction of the constraint. Top) the molecular function in question exhibits no constraint on the evolutionary rates of either functional domain class. Bottom) The molecular function in question can exhibit either a constraint on the evolutionary rate of a single class of domains (Restricted Constraint), or constrain both/all classes of domains (Global Constraint). In the case of a global constraint, arrows indicate the direction that the constraint is acting (i.e., the function of the Red domain directly constrains the Red domain and indirectly the Blue domain [Red → Blue]). *(B)* Each red circles indicate a protein domain in a single class of functions, the number of lines indicate the number of interactions in which each domain is involved. Each column models what would happen to the number of interactions as the number of domains increases from one to three under three different hypotheses. Left Column) No correlation—the total number of interactions that a protein is involved in does not change as the number of domains increases from one to three, Middle column) Direct correlation—the total number of interactions increases linearly as the number of domains increases from one to three, Right Column) Inverse correlation—the number of interactions decreases as the number of domains increase (increased specificity).

Evolutionary Rate Calculation

Each set of full length CDS and all sets of orthologous domains were aligned codon by codon using the ClutsalOmegacommandLine (Cock et al. 2009). This alignment was used to create a Maximum Likelihood tree using the RaxML commandline package in Biopython (Stamatakis 2006; Talevich et al. 2012). The sequence alignments and phylogenetic trees were then used to calculate the evolutionary rate (dN/dS) using PAML codeml program (seqtype = 1, NSsites = [0], CodonFreq = 2, fix_alpha = 1, kappa = 4.54006, model = 0, RateAncestor = 0), which were generated under a neutral evolution model (Nei and Gojobori 1986; Yang 2007;

Talevich et al. 2012). When either the dN = 0 or the dS = 0 a value of 0.0001 was automatically assigned by the software; these cases were removed from our analyses, as the assigned value does not accurately reflect the evolutionary rate of the protein. The resulting dN/dS scores were used as the evolutionary rate for all analyses in this study.

Number of TF–TF Interactions

To determine the number of PPIs between 2 TFs (TF–TF interactions) for each TF in our dataset we mined the STRING v9.1 protein interaction database (Franceschini et al. 2013). Our search was restricted to TF-TF interactions with High-throughput experimental evidence, thereby reducing the bias that might be created for well-studied genes. In total we were able to find 80,155 TF–TF interactions for 1,661 TFs (mean = 70.2 interactions per TF), accounting for 2% of the total interactions in the entire STRING v9.1 database. This yielded the number of TF–TF interactions for each of the 1,661 genes for which the data was available.

Number of DNA-Binding Targets

The number of DNA-binding targets for each TF was calculated from wgEncodeRegTfbsClusteredV3 database, found at http://genome.ucsc.edu/cgi-bin/hgTrackUi?db=hg19&g=wgEncodeRegTfbsClusteredV3. The ENCODE TfbsV3 data set consists of 690 Chip-Seq experiments, encompassing 161 TF and 91 different cell lines, with all TFs having at least 2 replicates. The data collection for all experiments followed the ENCODE Guidelines for Experiments for ChIP-seq experiments (Consortium 2004; Gerstein et al. 2012; Wang et al. 2013).

As part of the Standard Protocol for Encode ChIP-seq experiments, every experiment was subject to two peak calling procedures (Consortium 2004), SPP (which determines peaks based on by the Signal Score ([ChIP signal enrichment]/[input DNA signal])) (Kharchenko et al. 2008), and PeakSeq (which determines peaks based on the expected false discovery rate) (Rozowsky et al. 2009). Data shown is from the SPP peak calls, which were shown to be consistent for peak calling (Zhang et al. 2009). This yielded the number of DNA binding targets for each of the 161 genes for which the data were available.

GWAS Association

In order to determine which genes were associated with disease phenotypes, we mined the NHGRI GWAS catalog NHGRI GWAS catalog (Welter et al. 2014). All TF genes associated with SNPs that reached genome wide significance (P-value < 10^{-8}) in at least 1 GWAS were recorded as being a GWAS associated TF, all other TF genes were recorded as not being GWAS associated.

Statistics

In total, we calculated eight variables for each of the 1,988 TF encoding gene (no. of TF–TF interactions, no. of DNA binding targets, dN/dS of full TF, dN/dS of PPI domains, dN/dS of DNA binding domains, no. of PPI domains, no. of DNA binding domains, and GWAS association). We tested for correlation among the variables as shown in supplemental table S1, Supplementary Material online. The dN/dS measurements, no. of TF–TF interactions and no. of DNA-binding targets were natural log transformed in all analyses in this study. Correlation and significance (P-value) between all sets of continuous data were performed using the linear model package in R. Correlations including either the no. of PID or no. of DBD were calculated using Spearman Rank Correlation, all other correlations were calculated using Pearson correlation. Students t-test was used to test differences between GWAS and non-GWAS groups of TFs in regard to the no. of TF–TF interactions, DNA-binding targets, or dN/dS values. Heterogeneity among correlations was calculated based on a one-tailed Fisher r-to-z transformation to assess the significance of the differences between two correlations.

Results

Evolutionary Rates of Complete TF Coding Regions

In general, we found that increasing pleiotropy in TFs did constrain gene evolution as TFs with more interactions had a lower evolutionary rates regardless of the molecular function. The evolutionary rates of the TFs (as assessed by dN/dS) were negatively correlated with both the number of TF–TF interactions (Pearson $r = -0.310$, P-value = 5e-36; fig. 2A) and the number of DNA-binding targets (Pearson $r = -0.199$, P-value = 0.014; fig. 2B). However, the correlation of dN/dS with TF-TF interactions were not significantly stronger than the correlation with the DNA binding targets (P-value = 0.0823). Additionally, we saw a significant positive correlation between the numbers of TF–TF interactions and the DNA binding targets (Pearson $r = 0.337$, P-value = 2e-05; fig. 2C). Therefore, it is clear that having more TF–TF interactions and DNA binding targets is associated with slower gene evolution, but the effect size varies with domain function.

Evolutionary Rates of TF Protein Domains

The number of TF–TF interactions affected the evolution of both PPI and DNA binding domains. In contrast, the number of DNA-binding targets per TF only affected the evolution of the DNA binding domains (fig. 2). The evolutionary rates of both the PPI and DNA-binding domains were negatively correlated with the number of TF–TF interactions (Pearson $r = -0.191$, P-value = 9e-7; fig. 3A, and P-value = 6e-06; fig. 3B, respectively). The number of DNA-binding targets, however, only constrained the evolution of the DNA-binding domains (Pearson $r = -0.271$, P-value = 0.045; fig. 3D), as no

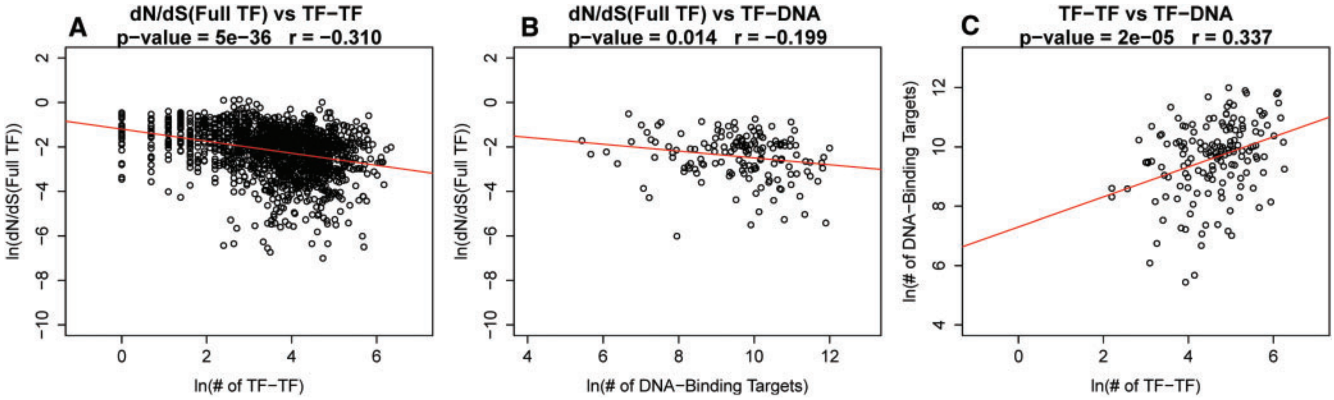

Fig. 2.—Pleiotropic constraints on protein evolutionary rate. X-axes represent the log transformed number of either DNA-binding targets or TF–TF interactions sand the Y axes are the log transformed evolutionary rates as measured by dN/dS for A and B. *(A)* The ln(evolutionary rate of the full TF) plotted against the ln(TF–TF interactions), *n* = 1,552. *(B)* The ln(evolutionary rate of the full TF) plotted against the ln(DNA-binding targets), *n* = 152. *(C)* The correlation between the log transformed number of TF–TF interactions and the log transformed number of DNA-binding targets, *n* = 154. dN/dS = evolutionary rate, TF–TF = no. of TF–TF interactions, DNA binding targets = no. of TF–DNA interactions.

significant correlation was seen with the evolutionary rates of domains involved in PPI (Pearson *r* = 0.081, *P*-value = 0.607; fig. 3C). These findings demonstrate that the specific pattern of constraint differs by domain function(s).

Number of Functional Domains Affects Gene Evolution

In addition to the number of binding partners (either PPI or protein-DNA) affecting gene evolution, the number of domains per TF positively correlated with the evolutionary rates of the TFs. We investigated how the number of PPI domains correlated to the evolutionary rates of these domains, as well as, to the number of TF–TF interactions. We found that as the number of PPI domains increased the evolutionary rates of the PPI domains also increased (Spearman *r* = 0.167, *P*-value = 1e-08; fig. 4A; table 1). This is consistent with a reduced constraint on domain evolution, as increasing numbers of PPI domains in a gene results in faster evolutionary rates.

Increasing numbers of PPI domains correlated with both fewer TF–TF interactions overall and per domain. The correlation between the number of PPI domains and the number of TF–TF interactions was highly significant and negative (Spearman *r* = −0.455, *P*-value = 1e-51; fig. 1B; table 1). Additionally, the number of PPI domains negatively correlated with the number of TF–TF interactions per PPI domain (Pearson *r* = −0.779, *P*-value = 2e-157; table 1; supplementary fig. S1*a*, Supplementary Material online). Therefore, overall more PPI domains correlated with fewer interactions per protein.

As the number of PIDs increased, the variance in the evolutionary rates of the PIDs decreased (Pearson *r* = −0.777, *P*-value = 0.0002; table 1; supplementary fig. S1*b*, Supplementary Material online). Additionally, as the number of PPI domains increased, the variance in the number of the TF–TF interactions also decreased (Pearson *r* = −0.592, *P*-value = 0.016; table 1; supplementary fig. S1*c*, Supplementary Material online). This relationship between the number of domains and both the domain evolutionary rates and number of TF–TF interactions was only observed for the PPI domains, as no significant correlations were found between the number of DNA binding domains and evolutionary rates of these domains nor with the number of DNA binding targets (table 1; supplemantary fig. S2, Supplementary Material online).

Pleiotropic Constraint in Disease Associating TFs

Finally, we tested whether mutations that predispose individuals to disease are more pleiotropic than those that do not, as such highly pleiotropic genes are more likely to have undergone selection. Such selection should constrain the evolutionary rates of more pleiotropic genes (Blekhman et al. 2008; Cai et al. 2009). We then compared the levels of pleiotropy between GWAS associated and non-GWAS associated TFs. TFs associated with human disease in a GWAS had significantly higher ln(no. of TF-TF interactions) than those not found in the GWAS catalog (GWAS associated genes (mean = 4.1), non-GWAS associated genes (mean = 3.5), *P*-value = 9e-14; fig. 5A). A similar trend was observed for the ln(no. of DNA binding targets); however, this did not reach statistical significant (GWAS associated genes (mean = 10), non-GWAS associated genes (mean = 9.5), *P*-value = 0.078; fig. 5B). The lack of statistical significance may be due to a greatly reduced sample size (*n* = 28). Additionally, GWAS-associated TFs are more evolutionarily constrained, as they appear to have lower ln(dN/dS) (GWAS associated genes (mean = −2.4), non-GWAS associated genes (mean = −2.1), *P*-value = 4e-6; fig. 5C). These results indicate that highly pleiotropic genes

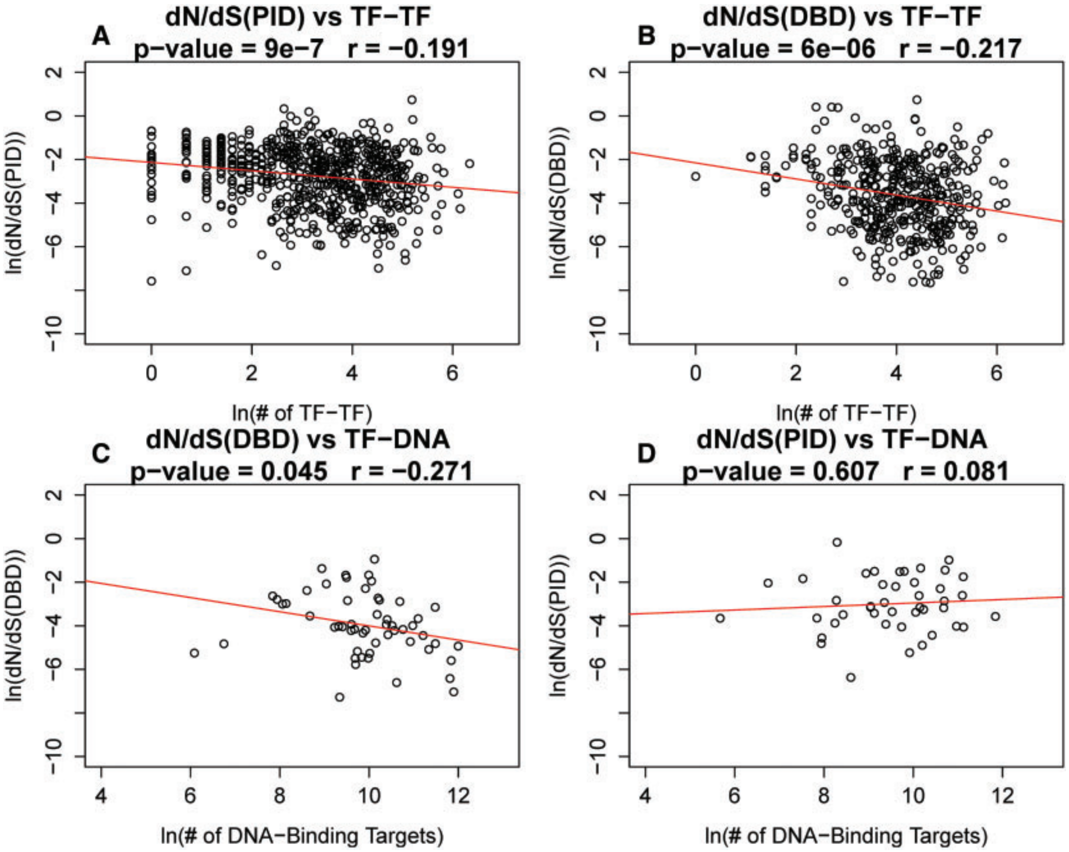

FIG. 3.—Pleiotropic constraint over functional domain evolutionary rate. X-axes represent the log transformed number of either DNA binding targets or TF–TF interactions and the Y axes are the log transformed evolutionary rates of functional domains, either protein– PID or DNA-binding Domain (DBD) as measured by dN/dS. *(A)* The ln(evolutionary rate of the PID) plotted against the ln(TF–TF interactions), $n = 651$. *(B)* The ln(evolutionary rate of the DBD) plotted against the ln(TF–TF), $n = 422$. *(C)* The ln(evolutionary rate of the PID) plotted against the ln(DNA-binding targets), $n = 44$. *(D)* The ln(evolutionary rate of the DBD) plotted against the ln(DNA-binding targets), $n = 55$. dN/dS, evolutionary rate; TF–TF, no. of TF–TF interactions; PID, protein-interacting domains; DBD, DNA-binding domains; DNA-binding targets, no. of TF–DNA interactions.

are more likely to be associated with a disease phenotype when mutated, and disease associations are reflective of evolutionary rates.

Discussion

Pleiotropic Constraints on Evolutionary Rates of the Entire TF CDS

Our study provided supportive evidence that pleiotropy constraints the evolution of TFs, but importantly also demonstrated that the constraints operate in a domain specific manner. TF–TF interactions constrain the evolutionary rates of the entire TF-CDSs. This is consistent with previous studies of PPI networks (Fraser et al. 2002, 2003; Fraser and Hirsh 2004; Fraser 2005; He and Zhang 2006; Kim et al. 2006; Salathe et al. 2006; Chang et al. 2013). However, unlike the prior studies we were also able to show that TF evolution was also constrained by the number of DNA binding targets. These data taken as a whole indicate that pleiotropic constraint is not limited to proteins with specific molecular functions but is a generic property of pleiotropy.

Constraints on Domain Specific Evolutionary Rates

We were also able to assess whether and the degree to which constraints operated within and across domains. TF–TF interactions not only constrained the entire gene CDS, as noted earlier, but the PPI domains also were significantly associated with constraint on the DNA-binding domains. In contrast, the constraint produced by the number of DNA-binding targets was restricted to the DNA binding domains, possibly explaining the overall weaker effect on the entire protein (fig. 6A). This reduced correlation could also result from the fact that many of the DNA motifs that a TF binds *in vitro* are not functional and hence would not affect the evolutionary rate of the protein. This is easy to argue as 60% of DNA-binding targets were not upstream of transcription start sites and therefore

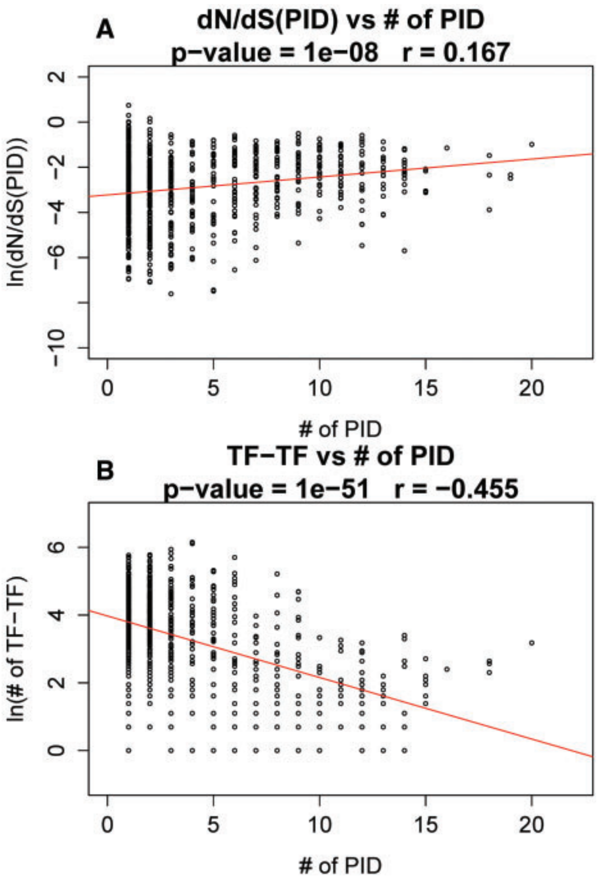

Fig. 4.—Pleiotropic constraint over functional domain count. X-axes for A–B show the number of PID identified within each TF. *(A)* The ln(evolutionary rate of the PID) plotted against the number of PID, *n* = 789. *(B)* The ln(TF–TF) plotted against the number of PID, *n* = 769. dN/dS, evolutionary rate; TF–TF, no. of TF–TF interactions; PID, protein-interacting domains; DBD, DNA-binding domains,.

are less likely to constrain evolution (Gerstein et al. 2012). Such noise in the data could lead to lower correlations between evolutionary rates and number of DNA-binding targets. However, if our lack of knowledge hides truly functional sites the degree of noise would be less.

Number of Domains Affect the Degree of Constraint

TFs with more PPI domains evolve at faster rates than those with fewer PPI domains, consistent with a reduced constraint on individual domains within each TF. This reduction in constraint can be explained by the finding that TFs with more PPI domains also tend to have fewer interactions per domain and per protein. Therefore, as each domain interacts with fewer proteins, there would be less constraint on each domain's evolution. This would result in faster evolution. However, these findings conflict with the prior literature, where it was shown that there was no significant correlation between the number PPI domains and average number of interactions in the human PPI network. Instead the correlations were driven by the proportion of each gene that encoded PPI domains (Xia et al. 2008). Additionally, in yeast it was shown that the more interacting surfaces that a protein has, the lower its evolutionary rate (Kim et al. 2006). While these studies do conflict with our results, it is important to distinguish these studies from our own, as they were not specifically looking at TFs. This distinction is important because of the overabundance of highly specific interactions and protein complexes found in the TF network (Thorsten and Valkhard 2014). It is therefore important to note that these findings (regarding the number of PPI domains) may not be reflective of other classes of genes outside of TFs; this possibility will require further analyses.

At first glance, the negative correlation between the number of PPI domains and the number of TF-TF interactions is counter-intuitive. However, this can be explained by the hypothesis that TFs have evolved greater numbers of PPI domains, to allow for higher specificity of domain interactions (fig. 6B). Consistent with this, it has been proposed that to increase the precision of gene regulation, protein interacting surfaces (mainly PPI domains) need to eliminate competition between prospective interacting partners; this can be accomplished by evolving domains specific to only a few interactions (Kim et al. 2006; Keskin and Nussinov 2007; Thorsten and Valkhard 2014). This may occur for one of several varieties of reasons; for example, if a TF is part of a larger complex of proteins you might expect it to possess multiple PPI domains, with each domain being designated to a specific interaction with another member of the complex. This would allow the components to tightly regulate each complex's assembly, stability, and function.

Robustness and Evolvability in TFs

Our results showed structural and functional elements of TF that significantly associate with their evolutionary rates. Our discussion has been largely focused on constraint, but such patterns amy also have implications for robustness and evolvability of genes. In a simplified view these three terms can be thought of as a spectrum; 1) constraint: the inability to tolerate mutation, 2) robustness: the ability to tolerate mutation without affecting function, and 3) evolvability: the ability to benefit from mutation (Masel and Trotter 2010) (resulting in negative, neutral, and positive selection, respectively). The largest implications derive from the correlations between the number of TF–TF interactions, the number of PID, and evolutionary rates. As argued earlier, TFs with many PIDs appear to have a higher specificity of interactions. The finding that these TFs evolve at faster rates also suggests that they are more robust, as mutations are less apt to disrupt their function(s) as a TF. Furthermore these mutations may allow the TF more likely to evolve novel interactions.

Table 1

Summary statistics from linear models between the values described in column 1 and either the no. of PID of the no. of DBD for each protein

	No. of PID			No. of DBD		
	No.	**Spearman *r***	***P*-value**	**No.**	**Spearman *r***	***P*-value**
ln(dN/dS of PIDs)	788	0.201	1.318E-08	220	−0.117	0.0834
Variance of ln(dN/dS of PIDs)	16	−0.777	1.07E-04	5	−0.584	0.335
ln(no. of TF–TF interactions)	768	−0.508	1.39E-51	683	−0.030	0.434
ln(no. of TF–TF interactions per PID)	768	−0.779	2.44E-157	223	−0.052	0.805
Variance of ln(no. of TF-TF interactions)	15	−0.592	0.016	6	−0.365	0.42
ln(dN/dS pf DBDs)	135	0.055	0.524	444	−0.036	0.448
Variance of ln(dN/dS of DBDs)	8	−0.144	0.734	6	−0.312	0.4959
ln(no. of DNA-binding targets)	57	−0.021	0.98	86	−0.084	0.439
ln(no. of DNA-binding targets per DBD)	33	−0.110	0.537	86	−0.234	0.029
Variance of ln(no. of DNA binding targets)	5	0.157	0.766	4	0.888	0.112

Note—For analyses of variance, each data point refers to the variance of either no. of TF–TF, no. of DNA-binding targets, or dN/dS for all TFs containing a specified number of domains (i.e., the variance of no. of TF–TF interactions for all TFs with 1 PID, 2 PID, 3 PID, etc.).

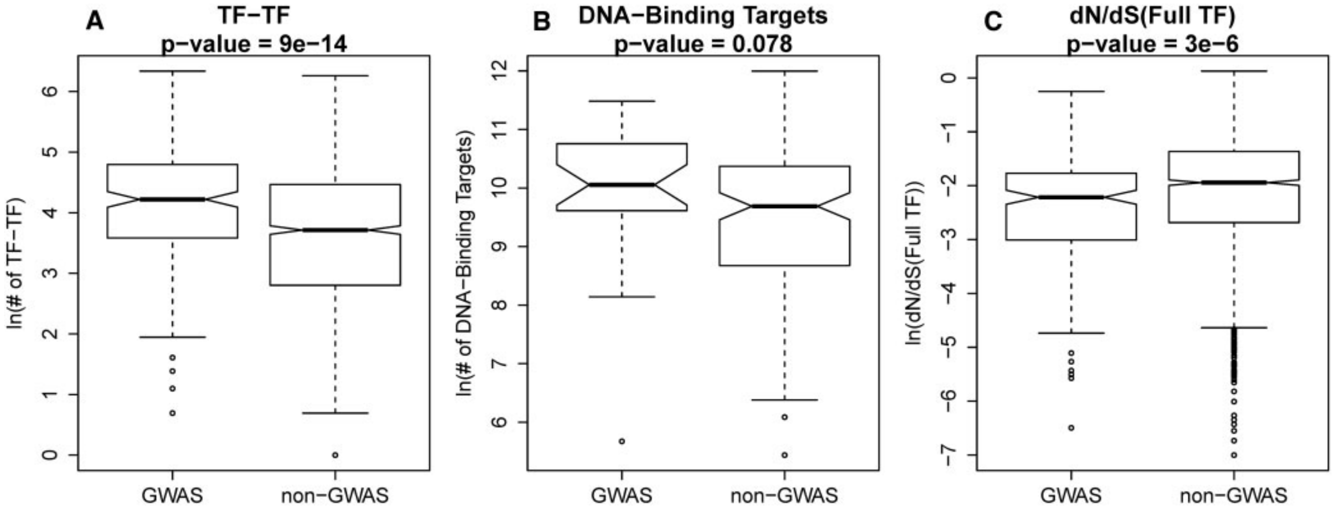

Fig. 5.—GWAS association with Pleiotropic functions. The X-axis of A–C splits the genes into two groups, genes that have been identified in at least one GWAS study and genes that have not (NO_GWAS). *(A)* The Y axis represents the log transformed number of TF–TF interactions, n of GWAS = 219, no of non-GWAS = 1,392. *(B)* The Y axis represents the log transformed DNA binding targets, no. of GWAS = 27, no. of non-GWAS = 134. *(C)* The Y axis represents the log transformed evolutionary rate of the full TF, no of GWAS = 223, no of non-GWAS = 1,708. Notches show 95% CI.

Correlation of the TF–TF and DNA-Binding Targets

We found a positive correlation between the number of TF–TF interactions and the number of DNA binding targets. It makes intuitive sense that these numbers are related to one another and suggests that there may be a causal relationship between them. This relationship could be due to either of the following two hypotheses:

1) As a TF acquires more TF–TF interactions over the course of evolution, these novel interactions facilitate a given TF's binding to more sites in the genome;

2) As a TF acquires more DNA binding targets over the course of evolution, these novel binding sites can place the TF in physical proximity to other TFs, facilitating more and novel TF–TF interactions.

Although speculative, we argue that the differential constraint on domain evolutionary rates between the TF–TF interactions and DNA-binding targets, shown in figure 3, indirectly supports hypothesis 1. The reason for this is that if the DNA binding targets are dependent on the TF–TF interactions, it is likely that the evolutionary rates of the DNA-binding domains would be correlated to the number of TF–TF interactions (fig. 3*C*). However, if the alternative were true, then we would expect to find that the number of DNA binding targets would significantly correlate with the evolutionary rates of the

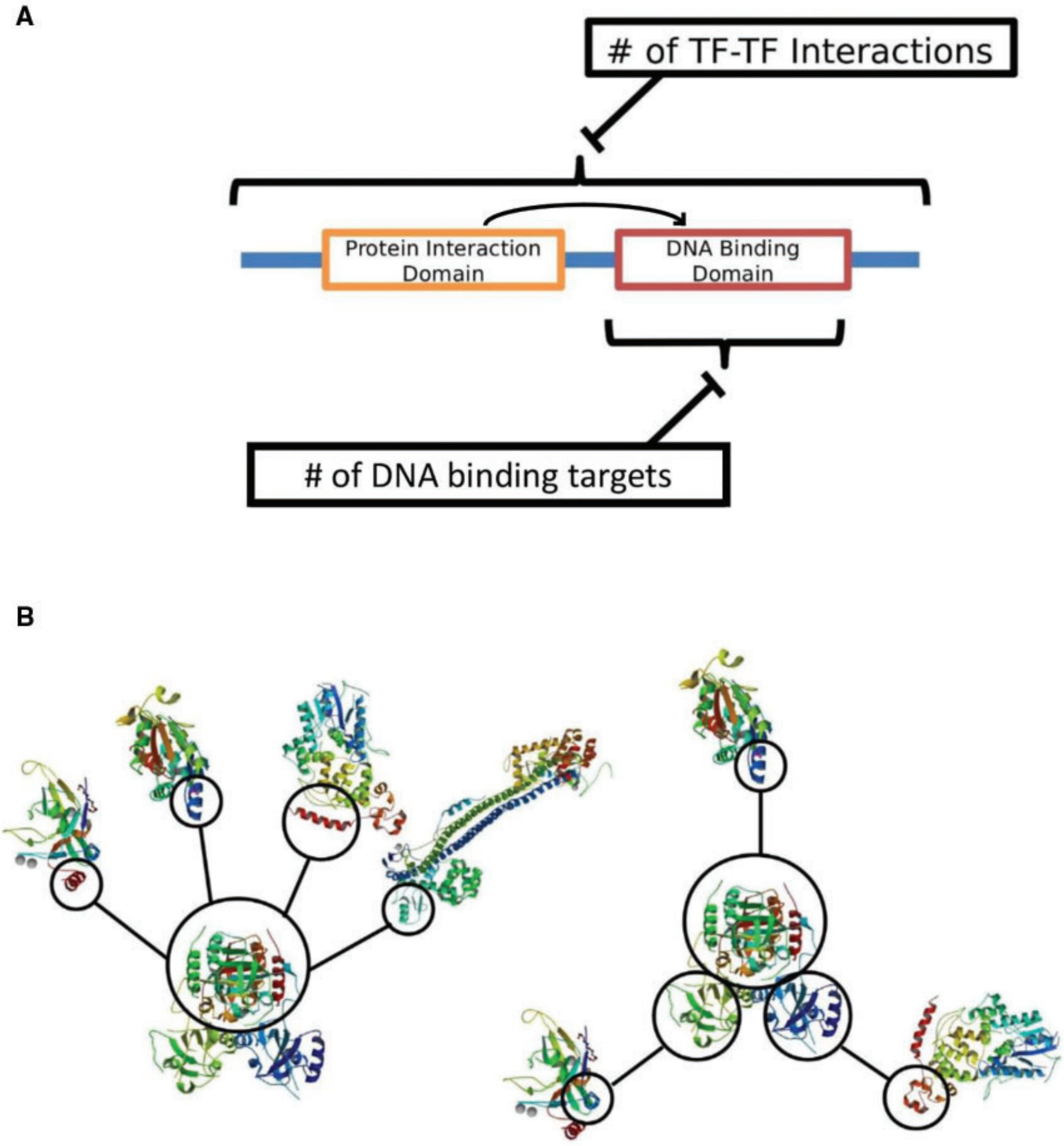

Fig. 6.—Possible molecular models of constraints. *(A)* Differential constrains applied to gene evolution from both the no. of TF–TF interactions and the no. of DNA binding targets. Blue line represents a linear diagram of a hypothetical protein containing 1 protein-interacting domain (Orange) and 1 DNA-binding domains (Red). *(B)* Ribbon diagram of a protein illustrating the effect of having a single protein-interacting domain (Left), versus multiple protein-interacting domains (right). Arrow shows direction of the global constraint.

PPI domains; since this was not observed this argues against alternative 2. It is also possible that this differential constraint could be the result of an unknown cofounder. Additionally, a similar model to hypothesis 1 has been put forth. In this model, as a transcriptional regulator (bound to DNA) interacts with new protein partners, the neighboring DNA sequences can then undergo selection to optimize the recruitment and affinity of the new protein interacting partner to the DNA sites (Tuch et al. 2008). This model illustrates how the gain of novel TF-TF interactions can facilitate new DNA binding targets (Tuch et al. 2008). This model also provides a more parsimonious means for evolving transcriptional circuitry, than hypothesis 2, due to the need for compensatory evolution. In hypothesis 1 this compensatory evolution would occur by mutating the neighboring DNA sequences; however, under hypothesis 2 the compensatory evolution would involve mutating the CDS in order to optimize the new TF–TF interactions. Because DNA is more robust than proteins, mutating DNA is likely to have less detrimental effects than mutating proteins. Furthermore, evolving new DNA binding targets would likely only facilitate a small number of novel TF–TF interactions, whereas new TF–TF interactions could potentially facilitate many DNA binding targets (one for each location where the new TF interacting partner is located on the genome).

Pleiotropy and Epistasis

The correlation between the number of TF–TF interactions and the number of DNA binding targets also raises an interesting question regarding the likely connection between pleiotropy and epistasis. At the molecular level, pleiotropy is defined as the ability a gene product to be involved in multiple molecular functions, and epistasis can be defined as the ability of one gene to modulate the function of another gene. Over evolutionary time newly acquired molecular functions (i.e., pleiotropy) are capable of altering the function of another gene (i.e., epistasis), the altered gene function can, in turn, facilitate the acquisition of additional novel functions (i.e. pleiotropy). A similar non-evolutionary argument has bene made for the relationship of epistasis and pleiotropy (Tyler et al. 2009).

Pleiotropic Constraint on Disease Genes

We showed that highly pleiotropic genes are more likely to be associated with a disease phenotype. This may be explained by two ideas: 1) The more processes a gene is involved in the more likely that gene is to be involved in an essential process, the disruption of which would result in a disease phenotype or 2) Disease phenotypes may arise from the disruption of multiple molecular processes, such as the disruption of cell cycle signaling pathways and apoptotic pathways in cancer progression (Vogelstein and Kinzler 2004; Gundem et al. 2015). These alternatives are not necessarily mutually exclusive. We also showed that disease associated genes have a high probability of being evolutionarily constrained, as they appear to present with significantly lower evolutionary rates than non-disease associated genes. This indicates that disease status does appear to reflect purifying selection, as the lower evolutionary rate is consistent with some alleles being removed from a population over evolutionary time (Blekhman et al. 2008; Cai et al. 2009). When both the pleiotropic and evolutionary data are taken together these findings show that this pleiotropic constraint does not only exist at the molecular level, but is also applicable at the phenotypic level (i.e., clinical phenotypes) (Sivakumaran et al. 2011; Wagner and Zhang 2011).

Limitations of This Study

There has been discussion as to how the bias of gene expression can influence the number of PPI that are known for a given protein. This is primarily driven by the fact that it is easier to identify interacting partners for highly expressed genes due to technological limitations. Several models have been put forth arguing that gene expression alone could affect evolutionary rates, regardless of pleiotropy. And while there is evidence to suggest that gene expression alone may constrain gene evolution, there have still been several studies that still show a significant (albeit weaker) correlation between the functional importance of a gene and evolutionary rates when controlled for gene expression (Wall et al. 2005; Zhang and He 2005). Unfortunately, there are several problems with interpreting these results in human gene evolution that are absent in model systems such as yeast, where much of this work has been done. Gene expression for complex animals are highly context dependent and the profiles will vary based on numerous factors, including which tissue the samples were taken from, race/ethnicity, age of the donor, and disease that the donor suffered from (such as from tumor cell lines). Therefore, adjustments for gene expression in human will likely change results and interpretation with respect to constraint and pleiotropy as a function of these and other factors. Nonetheless, we would argue that given the strength of some of the associations we detected the results for pleiotropy are likely to remain in at least some of the cases we observed.

Another limitation we encountered is how few TFs had DNA binding target data available. Due to the limited number of TFs with the available data, we found several correlations that were marginally significant or only near significant. With a small sample size it is difficult to say whether these correlations are real or not. Additionally, although not as big a problem, the PPI, domain descriptions and GWAS datasets are also incomplete, as not all interactions, domains, and disease associations are known for all TFs. With growing databases, the ability to assess the role of pleiotropy on gene evolution will improve and these relationships will continue to be resolved.

Conclusion

In this study, we showed that molecular pleiotropy is not only capable of constraining gene evolution, but that this constraint appears to differ by molecular function. We also showed that the number of PPI domains appears to uniquely constraint TFs, as the correlations we found were not observed for other sets of proteins in prior studies. These findings have implications for studies of protein and network evolution. In addition, TFs that are more pleiotropic are more likely to be associated with human disease risk. These findings emphasize the importance of studying pleiotropy to better understand how patterns of genetic variation are shaped and how pleiotropy can contribute to human diseases.

Supplementary Material

Supplementary table S1 and figures S1 and S2 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

This work was partially supported by grants T32GM008704 (K.N.C.) and P20GM103534 (S.M.W.) from the National Institutes of Health.

Literature Cited

Akashi H. 1994. Synonymous codon usage in drosophila-melanogaster-natural-selection and translational accuracy. Genetics 136:927–935.

Al-Maawali A, et al. 2015. Prenatal growth restriction, retinal dystrophy, diabetes insipidus and white matter disease: expanding the spectrum of PRPS1-related disorders. Eur J Hum Genet. 23:310–316.

Artieri CG, Haerty W, Singh RS. 2009. Ontogeny and phylogeny: molecular signatures of selection, constraint, and temporal pleiotropy in the development of Drosophila. BMC Biol. 7:42.

Blekhman R, et al. 2008. Natural Selection on Genes that Underlie Human Disease Susceptibility. Curr Biol. 18:883–889.

Cai JJ, Borenstein E, Chen R, Petrov DA. 2009. Similarly strong purifying selection acts on human disease genes of all evolutionary ages. Genome Biol Evol. 1:131–144.

Caspari E. 1952. Pleiotropic gene action. Evolution 6:1–18.

Chang X, Xu T, Li Y, Wang K. 2013. Dynamic modular architecture of protein-protein interaction networks beyond the dichotomy of 'date' and 'party' hubs. Sci Rep 3:1691.

Chuang TJ, Chiang TW. 2014. Impacts of pretranscriptional dna methylation, transcriptional transcription factor, and posttranscriptional microRNA regulations on protein evolutionary rate. Genome Biol Evol. 6:1530–1541.

Cock PJ, et al. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25:1422–1423.

Consortium EP. 2004. The ENCODE (ENCyclopedia Of DNA Elements) project. Science 306:636–640.

Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed proteins evolve slowly. Proc Natl Acad Sci U S A. 102:14338–14343.

Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134:341–352.

Franceschini A, et al. 2013. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res. 41:D808–D815.

Fraser HB. 2005. Modularity and evolutionary constraint on proteins. Nat Genet. 37:351–352.

Fraser HB, Hirsh AE. 2004. Evolutionary rate depends on number of protein-protein interactions independently of gene expression level. BMC Evol Biol. 4:13.

Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. 2002. Evolutionary rate in the protein interaction network. Science 296:750–752.

Fraser HB, Wall DP, Hirsh AE. 2003. A simple dependence between protein evolution rate and the number of protein-protein interactions. BMC Evol Biol 3

Gerstein MB, et al. 2012. Architecture of the human regulatory network derived from ENCODE data. Nature 489:91–100.

Gout JF, Kahn D, Duret L. Consortiu PP-G. 2010. the relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. Plos Genet. 6:

Gundem G, et al. 2015. The evolutionary history of lethal metastatic prostate cancer. Nature 520:353–357.

He X, Zhang J. 2006. Toward a molecular understanding of pleiotropy. Genetics 173:1885–1891.

Hodgkin J. 1998. Seven types of pleiotropy. Int J Dev Biol. 42:501–505.

Inagaki N, et al. 2006. Alpha B-crystallin mutation in dilated cardiomyopathy. Biochem Biophys Res Commun. 342:379–386.

Jordan IK, Wolf YI, Koonin EV. 2003. No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly. BMC Evol Biol. 3:1–8.

Keskin O, Nussinov R. 2007. Similar binding sites and different partners: implications to shared proteins in cellular pathways. Structure 15:341–354.

Kharchenko PV, Tolstorukov MY, Park PJ. 2008. Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat Biotechnol 26:1351–1359.

Kim PM, Lu LJ, Xia Y, Gerstein MB. 2006. Relating three-dimensional structures to protein networks provides evolutionary insights. Science 314:1938–1941.

Liu M, et al. 2006. Identification of a CRYAB mutation associated with autosomal dominant posterior polar cataract in a Chinese family. Invest Ophthalmol Vis Sci. 47:3461–3466.

Marchler-Bauer A, et al. 2011. CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res. 39: D225–D229.

Masel J, Trotter MV. 2010. Robustness and Evolvability. Trends Genet. 26:406–414.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol. 3:418–426.

Pal C, Papp B, Lercher MJ. 2006. An integrated view of protein evolution. Nat Rev Genet. 7:337–348.

Piasecka B, Lichocki P, Moretti S, Bergmann S, Robinson-Rechavi M. 2013. The hourglass and the early conservation models-co-existing patterns of developmental constraints in vertebrates. Plos Genet. 9.

Plate L. 1910. Vererbungslehre und deszendenztheorie. Festschrift Für Richard Hertwig 2:536–610.

Ran WQ, Kristensen DM, Koonin EV. 2014. Coupling between protein level selection and codon usage optimization in the evolution of bacteria and archaea. Mbio 5:e00956–14.

Ravasi T, et al. 2010. An atlas of combinatorial transcriptional regulation in mouse and man. Cell 140:744–752.

Roach JC, et al. 2007. Transcription factor expression in lipopolysaccharide-activated peripheral-blood-derived mononuclear cells. Proc Natl Acad Sci U S A. 104:16245–16250.

Rozowsky J, et al. 2009. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nat Biotechnol 27:66–75.

Salathe M, Ackermann M, Bonhoeffer S. 2006. The effect of multifunctionality on the rate of evolution in yeast. Mol Biol Evol. 23:721–722.

Sivakumaran S, et al. 2011. Abundant pleiotropy in human complex diseases and traits. Am J Hum Genet. 89:607–618.

Stamatakis A. 2006. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690.

Stearns FW. 2010. One hundred years of pleiotropy: a retrospective. Genetics 186:767–773.

Talevich E, Invergo BM, Cock PJ, Chapman BA. 2012. Bio.Phylo: a unified toolkit for processing, analyzing and visualizing phylogenetic trees in Biopython. BMC Bioinformatics 13:209.

Tang W, Zhang J, Lin D. 2014. Pleiotropic enrichment analysis with diverse omics data. Advance Genet Eng 3:1–2.

Thorsten W, Valkhard H. 2014. Identifying transcription factor complexes and their roles. Bioinformatics 30:415–421.

Tuch BB, Li H, Johnson AD. 2008. Evolution of eukaryotic transcription circuits. Science 319:1797–1799.

Tyler AL, Asselbergs FW, Williams SM, Moore JH. 2009. Shadows of complexity: what biological networks reveal about epistasis and pleiotropy. Bioessays 31:220–227.

Vankeerberghen A, Cuppens H, Cassiman JJ. 2002. The cystic fibrosis transmembrane conductance regulator: an intriguing protein with pleiotropic functions. J Cyst Fibros 1:13–29.

Vogelstein B, Kinzler KW. 2004. Cancer genes and the pathways they control. Nat Med 10:789–799.

Wagner GP, Zhang J. 2011. The pleiotropic structure of the genotype-phenotype map: the evolvability of complex organisms. Nat Rev Genet. 12:204–213.

Wall DP, et al. 2005. Functional genomic analysis of the rates of protein evolution. Proc Natl Acad Sci U S A. 102:5483–5488.

Wang J, et al. 2013. Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium. Nucleic Acids Res. 41:D171–D176.

Welter D, et al. 2014. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 42:D1001–D1006.

Williams GC. 1957. Pleiotropy, natural selection, and the evolution of senescence. Evolution 11:398–411.

Xia K, Fu Z, Hou L, Han JDJ. 2008. Impacts of protein–protein interaction domains on organism and network complexity. Genome Res. 18:1500–1508.

Yang JR, Chen XS, Zhang JZ. 2014. Codon-by-Codon Modulation of Translational Speed and Accuracy Via mRNA Folding. Plos Biol. 12.

Yang JR, Liao BY, Zhuang SM, Zhang JZ. 2012. Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. Proc Natl Acad Sci U S A. 109:E831–E840.

Yang JR, Zhuang SM, Zhang J. 2010. Impact of translational error-induced and error-free misfolding on the rate of protein evolution. Mol Syst Biol. 6:

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24:1586–1591.

Zhang JZ, He XL. 2005. Significant impact of protein dispensability on the instantaneous rate of protein evolution. Mol Biol Evol. 22:1147–1155.

Zhang JZ, Yang JR. 2015. Determinants of the rate of protein sequence evolution. Nat Rev Genet. 16:409–420.

Zhang X, et al. 2009. A myelopoiesis-associated regulatory intergenic noncoding RNA transcript within the human HOXA cluster. Blood 113:2526–2534.

Zhou T, Drummond DA, Wilke CO. 2008. Contact density affects protein evolutionary rate from bacteria to animals. J Mol Evol. 66:395–404.

Associate editor: Eric Bapteste